



*Enteropathogen Resource Integration Center*  
Bioinformatics Resource Center

# *Genome Annotation in ERIC*

Guy Plunkett III

University of Wisconsin-Madison

*BRC2 Meeting*

Virginia Tech

May 16, 2005



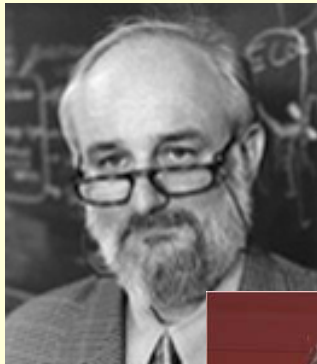
# ERIC Team:



## ERIC Curators:



Guy Plunkett III, Ph.D. - Senior Curator;  
David Bowen, Ph.D.; Val Burland, Ph.D.; Eric Cabot, Ph.D.;  
Jeremy Glasner, Ph.D.





# Enteropathogen Resource Integration Center

diarrheagenic *E. coli* genomes published and in progress

EHEC O157:H7 (2 genomes in ERIC)

EAgEC (Sanger Centre)

EPEC (Sanger Centre)

13 others, including

NIH-MGP (TIGR)

Universität Münster

Kitasato University

Institut Pasteur and Genoscope

*\*thanks to J. Kaper*

The screenshot shows a web browser window with the address bar displaying the URL: [http://www.ericbrc.org/eric/portal/media-type/html/user/anon/page/default.pml/template/Enteroaggregative\\_Escherichia\\_](http://www.ericbrc.org/eric/portal/media-type/html/user/anon/page/default.pml/template/Enteroaggregative_Escherichia_). The page title is "Enteroaggregative Escherichia coli (EAgEC)". The table below contains the following data:

Other names	N/A
Strain	O42
Serotype	O44:H18
Chromosome and Plasmid(s) Size	5,241,977 bp (chromosome); 113,346 bp (plasmid)
Genomic Sequence Status	Sequencing complete
Assembly Status	Final assembly at ~7.6X coverage; 1 plasmid
Sequencing Center	<a href="#">The Wellcome Trust Sanger Institute</a>
Sequencing Center Contact(s)	<a href="#">Julian Parkhill</a> or <a href="#">Bart Barrell</a>
Trace Files Status	Availability unknown - not online
Computational Annotation Status	Annotation ongoing by primary sequencing center
Manual Curation Status	Long term curation plans unknown
GenBank/EMBL/DBJ Entries	Not Available
NCBI RefSeq Entries	<a href="#">NC_004504</a> (BLAST only)

At the bottom of the page, there is a link: [For more information, please contact info@ERICBRC.org](#).



# Enteropathogen Resource Integration Center

*Shigella* genomes published and in progress

*S. flexneri* (2 genomes in ERIC)

*S. sonnei* (Sanger Centre)

*S. dysenteriae* (Sanger Centre)

others, including

*S. boydii* (TIGR)

*S. dysenteriae* (TIGR)

several from China?

ERIC BRC - Mozilla Firefox

http://www.ericbrc.org/eric/portal/media-type/html/user/anon/page/default.pml/template/AdhocReportScreen.vmcjsession

Enteropathogen Resource Integration Center  
Bioinformatics Resource Center

Pathogenic Organism	Isolate/Strain	Genomic Sequence Status	ERIC Genome Sequence Status
<b>Diarrhoeagenic <i>Escherichia coli</i></b>			
<a href="#">Enterohemorrhagic <i>Escherichia coli</i> (EHEC)</a>	EDL933 (ATCC 700927)	Sequencing complete	In ERIC BRC
<a href="#">Enterohemorrhagic <i>Escherichia coli</i> (EHEC)</a>	Sakai (RIMD 0509952)	Sequencing complete	In ERIC BRC
<a href="#">Enterohaemorrhagic <i>Escherichia coli</i> (EHEC)</a>	042	Sequencing complete	Not Yet Available
<a href="#">Enteropathogenic <i>Escherichia coli</i> (EPEC)</a>	E2349/69	Finishing/gap closure	Not Yet Available
<b><i>Shigella</i> spp.</b>			
<a href="#">Shigella flexneri</a>	2457T (ATCC 700930)	Sequencing complete	In ERIC BRC
<a href="#">Shigella flexneri</a>	301	Sequencing complete	In ERIC BRC
<a href="#">Shigella sonnei</a>	S3G	Finishing/gap closure	Not Yet Available
<a href="#">Shigella dysenteriae</a>	M131649 (M131)	Finishing/gap closure	Not Yet Available
<b><i>Salmonella</i> spp.</b>			
<a href="#">Salmonella Typhi</a>	Ty2 (ATCC 700931)	Sequencing complete	In ERIC BRC
<a href="#">Salmonella Typhi</a>	CT19	Sequencing complete	In ERIC BRC
<a href="#">Salmonella Typhimurium</a>	LT2 (ATCC 700720)	Sequencing complete	In ERIC BRC
<a href="#">Salmonella Typhimurium</a>	DT104	Finishing/gap closure	Not Yet Available
<a href="#">Salmonella Typhimurium</a>	SL1344	Finishing/gap closure	Not Yet Available
<a href="#">Salmonella Enteritidis</a>	PT4	Sequencing complete	Not Yet Available
<a href="#">Salmonella bongori</a>	12419	Sequencing complete	Not Yet Available
<a href="#">Salmonella Paratyphi A</a>	ATCC 9150	Sequencing complete	(in queue)



# Enteropathogen Resource Integration Center

*Salmonella* genomes published and in progress

Typhi (2 genomes in ERIC)

Typhimurium (1 genome in ERIC)

Paratyphi A (1 genome in ERIC)

Choleraesuis (1 genome in ERIC)

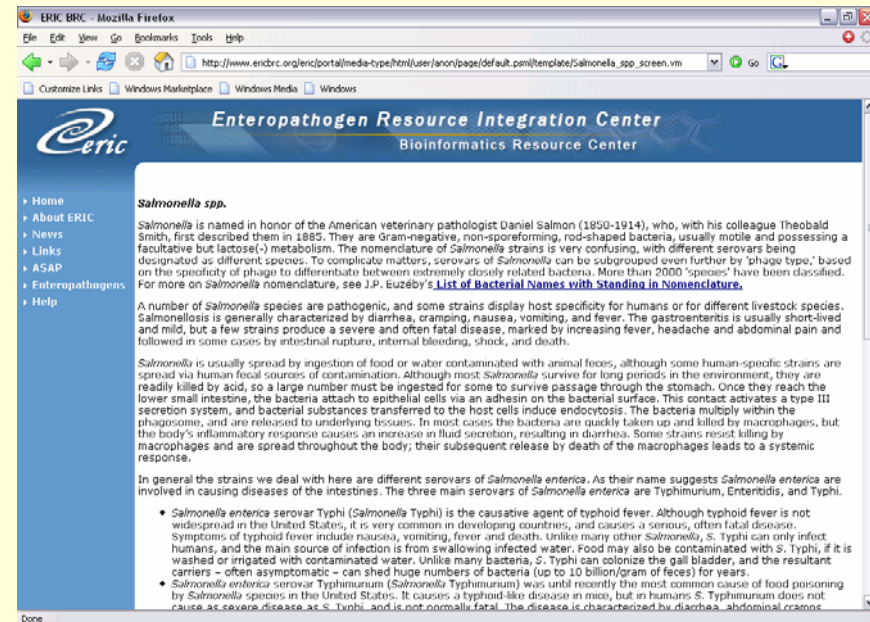
12 others, including

Sanger Centre

Washington University

University of Illinois

*\*salmonella.org*





# Enteropathogen Resource Integration Center

*Yersinia* genomes published and in progress

*Y. pestis* (3 genomes in ERIC)

*Y. enterocolitica* (Sanger Centre)

10 - 12 others including  
TIGR

ERIC BRC - Mozilla Firefox

http://www.ericbrc.org/eric/portal/media-type/html/user/anonymous/page/default.jsp?templateId=1001&reportScreen=vmjsession

<a href="#">Salmonella Enteritidis</a>	PT4	Sequencing complete	Not Yet Available
<a href="#">Salmonella bongori</a>	12419	Sequencing complete	Not Yet Available
<a href="#">Salmonella Paratyphi A</a>	ATCC 9150	Sequencing complete	(in queue)
<a href="#">Salmonella Paratyphi B</a>	SPB7	Survey shotgun sequencing (4X)	Not Yet Available
<a href="#">Salmonella Arizonae</a>	RSK2980 (246-86)	Survey shotgun sequencing (4X)	Not Yet Available
<a href="#">Salmonella Diarizonae</a>	CDC 01-0005 (ATCC BAA-639)	Survey shotgun sequencing (4X)	Not Yet Available
<a href="#">Salmonella Enteritidis</a>	LK5	Finishing/gap closure	Not Yet Available
<a href="#">Salmonella Paratyphi C</a>	[Not Indicated]	Finishing/gap closure	Not Yet Available
<a href="#">Salmonella Pullorum</a>	[Not Indicated]	Finishing/gap closure	Not Yet Available
<a href="#">Salmonella Dublin</a>	[Not Indicated]	Finishing/gap closure	Not Yet Available
<a href="#">Salmonella Choleraesuis</a>	[Not Indicated]	Sequencing complete	Not Yet Available
<a href="#">Salmonella Gallinarum</a>	287/91	Finishing/gap closure	Not Yet Available
<b><a href="#">Yersinia enterocolitica</a></b>			
<a href="#">Yersinia enterocolitica</a>	8081 (NCTC 13174)	Sequencing complete	Not Yet Available
<b><a href="#">Yersinia pestis</a></b>			
<a href="#">Yersinia pestis biovar Mediaevalis</a>	KIM	Sequencing complete	In ERIC BRC
<a href="#">Yersinia pestis biovar Mediaevalis</a>	91001 - avirulent to humans	Sequencing complete	In ERIC BRC
<a href="#">Yersinia pestis biovar Orientalis</a>	CO92	Sequencing complete	In ERIC BRC

For more information, please contact [info@ERICBRC.org](mailto:info@ERICBRC.org)

Copyright © 2004-2005 SRA International, Inc.  
Site updated March 09, 2005



## Enteropathogen Resource Integration Center

in addition to whole genomes, other relevant sequences that might be added include:

plasmids, bacteriophage, “pathogenicity islands,” and other mobile genetic elements

e.g., pWR501 (*S. flexneri* 5A), R27 (*Salmonella* Typhi)

contigs from genome-scanning projects





## Enteropathogen Resource Integration Center

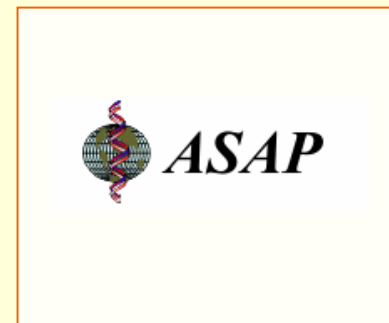
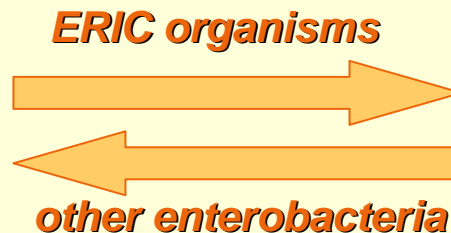
related genomes outside the BRC set are mirrored from UW for comparative purposes:

commensal *E. coli* (K-12, others in progress)

extraintestinal *E. coli* (CFT073, others in progress)

*Y. pseudotuberculosis*

... and others as they become available

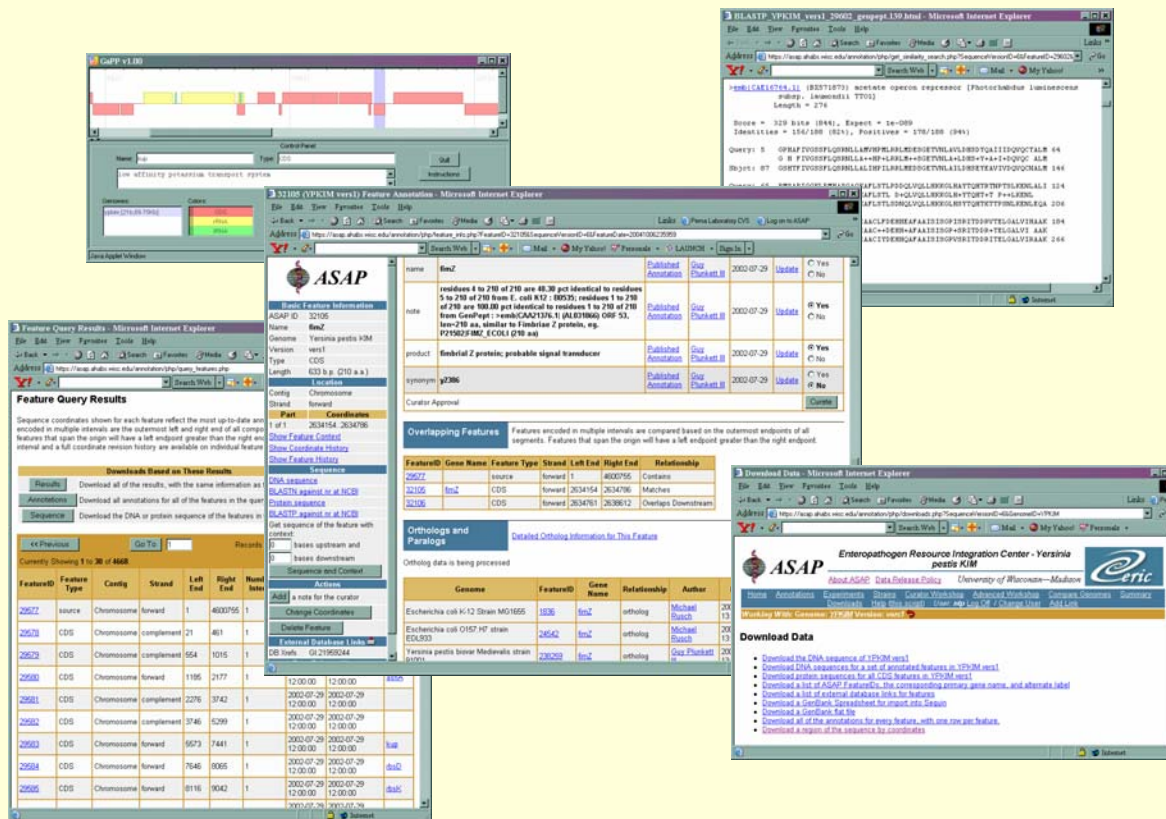






# Enteropathogen Resource Integration Center

## Bioinformatics Resource Center



**ASAP (A Systematic Annotation Package) for community annotation of genomes meets needs for direct, community annotation (with authorship and annotation history tracking), multiple annotations of features, evidence codes, using controlled vocabularies, curatorial review, support of cross-genome comparisons, and web-based updating and access.**



# Enteropathogen Resource Integration Center

## ERIC-BRC Annotation

ASAP can be populated with:

- Unannotated draft genome sequences
- Preliminary annotations in tabular format and associated sequences in FASTA format
- Genbank flatfiles
- For the *E. coli* and *Shigella spp.* being sequenced at TIGR, plan to populate ASAP with annotations from the TIGR automated analysis pipeline
- Can track versions of sequences and versions of annotations



# Enteropathogen Resource Integration Center

## ERIC-BRC Annotation

- All standard INSD Feature types are supported, plus additional custom types: conserved segment, contig, delimiter (for separating contigs in a “pseudomolecule,” island, prophage)
- For each Feature, a detailed annotation page supports standard and custom annotations (aka “qualifiers”)
- Basic information: type of Feature (e.g., CDS, rRNA, tRNA), version of the genome, location within the genome (coordinates and orientation), actual sequences (nt, aa, flanking)
- Rich Annotations: comments, function including Gene Ontology (GO), molecular interaction, mutant phenotypes, structures in PDB

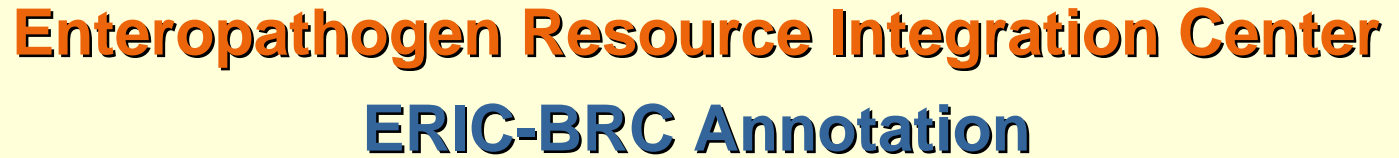
**ASAP**  
Basic Feature Information  
ASAP ID: 99006  
Name: sptP  
Genome: Salmonella typhimurium LT2  
Version: v1  
Type: CDS  
Length: 1832 b.p. (543 a.a.)

**Location**  
Contig: complement  
Strand: complement  
Part: Coordinates  
1 of 1: 3022071..3023702  
[Show Feature Content](#)  
[Show Coordinate History](#)  
[Show Feature History](#)

**Sequence**  
DNA sequence  
BLASTN against nr at NCBI  
Protein sequence  
BLASTP against nr at NCBI  
Get sequence of the feature with context:  
0 bases upstream and  
0 bases downstream  
[Sequence and Context](#)

**Actions**  
[Add a note for the curator](#)  
[Change Coordinates](#)  
Done

	Annotation	Value	Date	Status
comment	By temporal regulation, SptP antagonizes the host membrane ruffling induced by translocated effector SopE upon invasion. Persists longer in the host cell than SopE due to slower kinetics of host proteasome-dependent degradation. Half-lives of both proteins are encoded in their N-terminal secretion/translocation signal regions. SptP promotes host recovery to normal architecture despite infection, benefiting both host and pathogen.	Experimental - this species or strain	2005-01-25	Uncurated
function	secreted effector protein	Experimental - this species or strain	2005-01-24	Uncurated
function	Reverses host proinflammatory invasion response by tyrosine phosphatase action on host MAP kinase Erk signalling	Experimental - this species or strain	2005-01-24	Uncurated
function	Translocated into host cells.	Experimental - this species or strain	2005-01-24	Uncurated
function	Reverses invasion-induced actin rearrangements in host cells, by GTPase-activation of host Rac-1 and Cdc42.	Experimental - this species or strain	2005-01-24	Uncurated
GO biological process	GO:0030254 (GO:0030254 type III protein secretion system)	Experimental - this species or strain	2005-01-05	Uncurated
GO biological process	GO:0009405 (GO:0009405 pathogenesis)	Experimental - this species or strain	2005-01-21	Uncurated
GO molecular function	GO:0005096 (GO:0005096 GTPase activator activity)	Experimental - this species or strain	2005-01-05	Uncurated
GO molecular function	GO:0004725 (GO:0004725 protein tyrosine phosphatase activity)	Experimental - this species or strain	2005-01-05	Uncurated
locus tag	STM2878	Published Annotation	2004-11-12	Approved
	Interacts with host intermediate filament protein vimentin.	Experimental -		







# Enteropathogen Resource Integration Center

## ERIC-BRC Annotation

### GOALS

- up-to-date core annotations for existing and new genomes
- standardization across genomes where appropriate
- value added through
  - annotation/review by 5 dedicated curators
  - community contributions

### STRATEGIES

- leverage similarities across genomes
- focus on genes and processes relevant to BRC goals



# Enteropathogen Resource Integration Center

## ERIC-BRC Annotation

### TASK BASED APPROACH TO ANNOTATION

Tasks are motivated by:

- relevance to virulence/pathogenicity
- community interest
- urgent need to standardize features across genomes

Advantages:

- curators develop and apply expertise in focused areas
- genes with related functions are handled together
- related genes from different genomes are handled together



# Enteropathogen Resource Integration Center

## ERIC-BRC Annotation

### Attachment and colonization:

- Fimbriae and adhesins
- Quorum sensing
- Tissue tropism
- Ability to use specific nutrients localized in the target zone
- Resistance to host defense (acid resistance; urease, serum resistance)
- Drug resistance mechanisms (efflux pumps)

### Invasion of host cells:

- Effectors translocated into the host cell
- Cause changes to host cell morphology
- Persistence in host cell
- Resistance to or evasion of host defense mechanisms (oxidation, phagocytosis)
- Interaction with internal host cell membranes (vacuoles, Golgi)
- Bacterial motility in host cells
- Replication in host cells
- Mechanisms of dissemination to blood and other tissues
- Suppression of host immune response
- Regulation of invasion systems: bacterial sensor/effector systems, inducible gene expression

### Secretion systems:

- Types I – IV mechanisms and components, specialization
- Chaperones





# Enteropathogen Resource Integration Center

## ERIC-BRC Annotation

### Toxins:

- Heat-stable enterotoxins

- Heat-labile enterotoxins

- AB-type toxins

- RTX-type

- Hemolysins

- Tc-type

- Toxin delivery mechanisms

### Serotypes and surface antigens:

- Colanic acid

- O-antigen

- Enterobacterial Common Antigen (ECA)

- LPS

- H- and K-antigens

Islands, phages, plasmids and mobile DNA

Drug resistance

Iron acquisition and storage

Aerobic/Anaerobic pathways

Fatty acid and polyketide synthesis: multimodular (“factory”) proteins

The genetic basis for diagnostic tests in current use

Non-coding small RNAs



# Enteropathogen Resource Integration Center

## ERIC-BRC Annotation

### Start Site Reassessment

- whole-genome proteomics studies and gene-based microarray designs assume accurate assessment of coding sequences
- even the best gene prediction tools have problems with genes acquired by horizontal transfer and those that encode small proteins or proteins with transmembrane domains and/or signal peptides.
- such genes include some of the best candidates for diagnostics (pathogen-specific genes) and vaccine development (e.g., surface-exposed and exported proteins).
- wrong assignments in other genomes are readily propagated to new genomes during annotation.
- identifying coding sequences and annotating the correct translation start, especially in the absence of experimental evidence, relies on a variety of tools, including comparison across ortholog sets, codon usage patterns, RBS (SD) predictions, and even transcription start predictions



# Enteropathogen Resource Integration Center

## ERIC-BRC Annotation

### Pseudogenes

- mutant genes in a genome sequence, usually in comparison to a related genome where the wild-type or “ancestral” state is seen
- distinguished from missense mutations, where the gene is still intact but may have altered functionality
- genes disrupted by in-frame stop codons, frameshifts, insertion of IS elements, prophages, or islands; gene remnants, the aftermath of deletions, rearrangements, etc.
- implicated in the evolution of pathogens and their adaptation to new niches.
- inconsistently annotated in current genomes, and can actually be thought of in two conflicting ways:
  - in the context of evolution and comparative genomics: is a given gene present or absent in a particular genome, and is it intact or disrupted?
  - in the context of a given pathogen, are any partial genes actually translated? is the consequence of a pseudogene something other than the straightforward loss of function?

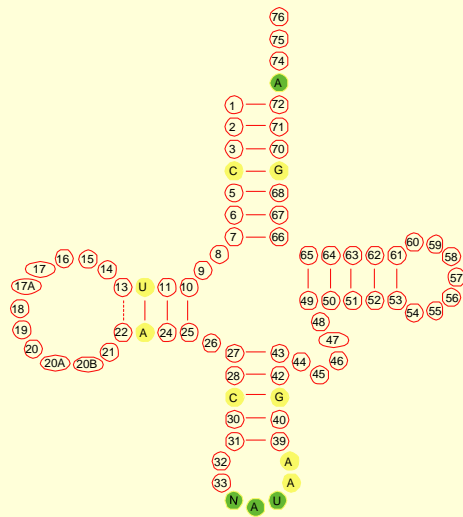


# Enteropathogen Resource Integration Center

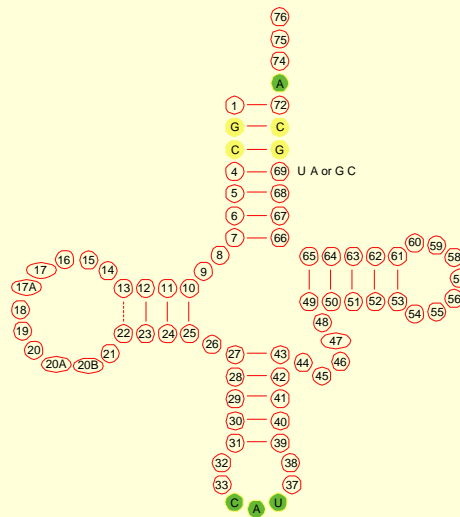
## ERIC-BRC Annotation

### Fine-tuning tRNA annotations

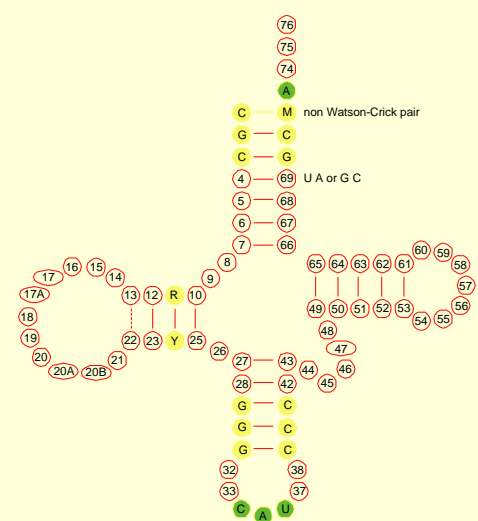
- Islands often integrated into or near tRNA genes
- tRNA genes within islands: improved expression of horizontally acquired genes?
- tRNAscan-SE does an excellent job, but does not distinguish between different tRNAs with CAT anticodons:



tRNA-Ile



tRNA-Met  
(elongator)

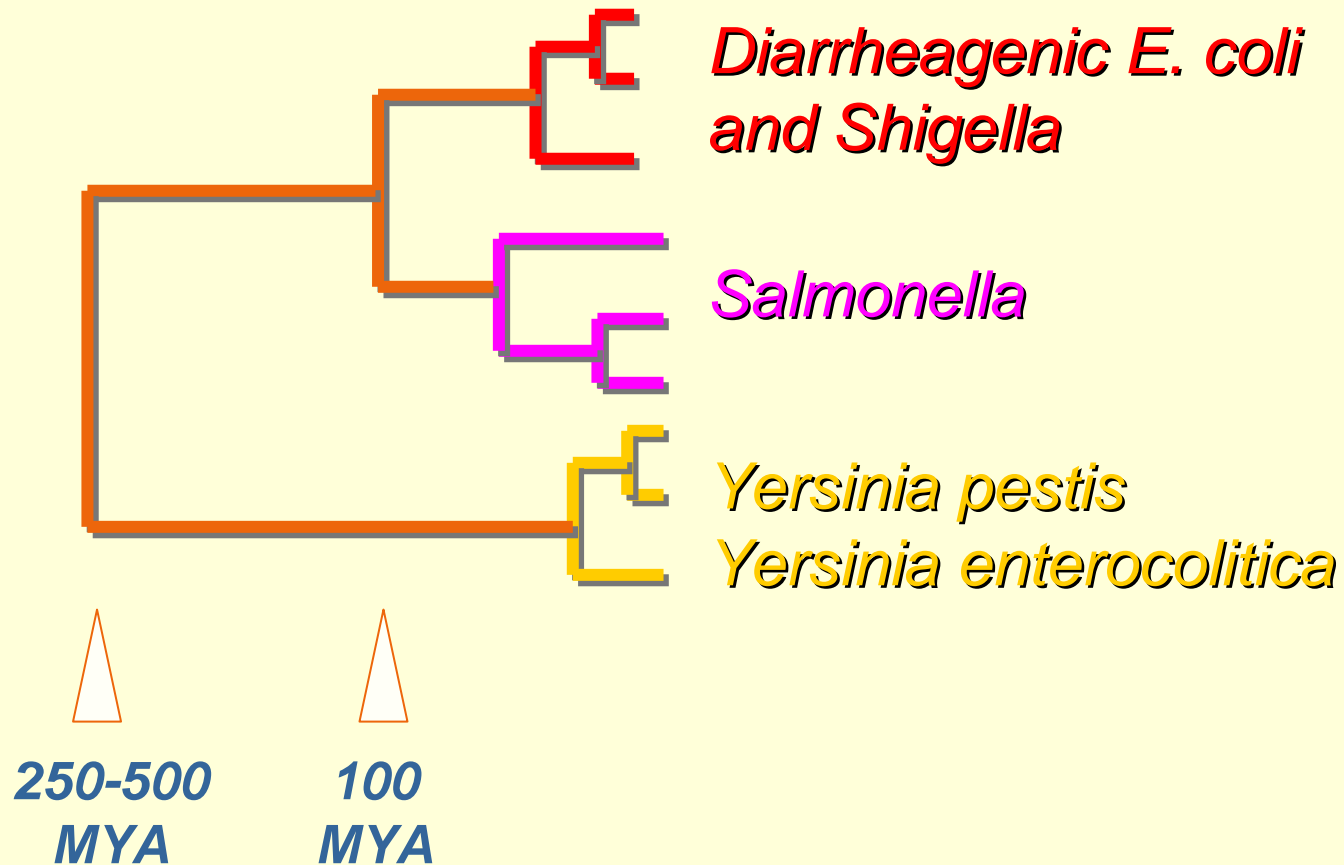


tRNA-fMet  
(initiator)



# Enteropathogen Resource Integration Center

## ERIC-BRC Annotation





# Enteropathogen Resource Integration Center

## ERIC-BRC Annotation

### Orthologs

- Semi-automated ortholog determination using filtered BLASTP reciprocal best hits
- Reviewed, corrected and augmented by curators
- Annotators/curators can transfer annotations from one ortholog to another
- Lists of shared and unique genes to query and download.

The screenshot displays the ERIC-BRC Annotation web interface. The top page, 'Show Orthologs for a Sequence', shows the 'Enteropathogen Resource Integration Center - Escherichia coli O157:H7 strain EDL933' and a list of orthologs. The bottom page, 'Feature Orthologs', shows a table of orthologs for the feature 'ECOL833 Feature 26754'.

Type	Data	A7	Version	Copy	Type	Data	A7	Version
codon start	1	A	NCBI	..	db ref	GI12518449	A	NCBI
db ref	GI12336836	A	NCBI	..	function	Factor: Extracellular function: Secreted protein	A	NCBI
evidence	inf_expansion	A	NCBI	..	status tag	Z5162	A	NCBI
feature tag	ECOL833	A	NCBI	..	name	ECOL833	A	NCBI
note	translocated into receptor 10 (0027) [Escherichia coli O157:H7 strain EDL933]	A	NCBI	..	product	protein to associated intracellular receptor protein	A	NCBI
product	translocated intracellular receptor 10	A	NCBI	..	synonym	Z5162	A	NCBI
protein	BARC1706A.1	A	NCBI	..				



# Enteropathogen Resource Integration Center

## ERIC-BRC Annotation

Mauve multiple alignments, with display of annotated features

